Penetrance rate estimation in autosomal dominant conditions

Paulo A. Otto[1] and Andréa R.V.R. Horimoto[2]

[1]*Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil.*
[2]*Laboratório de Genética e Cardiologia Molecular, Faculdade de Medicina, Instituto do Coração, Universidade de São Paulo, São Paulo, SP, Brazil.*

Abstract

Accurate estimates of the penetrance rate of autosomal dominant conditions are important, among other issues, for optimizing recurrence risks in genetic counseling. The present work on penetrance rate estimation from pedigree data considers the following situations: 1) estimation of the penetrance rate K (brief review of the method); 2) construction of exact credible intervals for K estimates; 3) specificity and heterogeneity issues; 4) penetrance rate estimates obtained through molecular testing of families; 5) lack of information about the phenotype of the pedigree generator; 6) genealogies containing grouped parent-offspring information; 7) ascertainment issues responsible for the inflation of K estimates.

Key words: penetrance rate, maximum likelihood method, recurrence risks, genetic counseling.

Introduction

Human autosomal dominant diseases are extremely rare conditions in which affected individuals are heterozygotes. Many of these heterozygous genotypes exhibit the phenomenon of incomplete penetrance. For this set of rare conditions the penetrance rate is therefore understood as the probability of a heterozygote presenting the disease (or, at least, presenting a minimum number of signs and symptoms that enable his/her identification as a carrier of the deleterious allele). Other details, as well as a full review of the subject can be found in Horimoto and Otto (2008). Accurate estimates of the penetrance value K are important not only for determining genetic disease risks in families with segregating cases of autosomal dominant disorders, but also for performing linkage studies. Crude penetrance estimates can be derived by dividing the observed number of diseased (penetrant) individuals by the number of obligate carriers (penetrant as well as obligate non-penetrant, that is, normal individuals with several affected offspring or normal individuals with affected parent and child). Presently the penetrance parameter can be estimated on a routine basis by computer programs that perform segregation analysis or the estimation of linkage based on complex pedigree structures that cannot be expressed in closed form, such as the classical S.A.G.E. (S.A.G.E., 2009) and LINKAGE (Lathrop *et al.*, 1985) programs.

Send correspondence to Paulo A. Otto. Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, Caixa Postal 11461, 05422-970 São Paulo, SP, Brazil. E-mail: otto@usp.br.

Rogatko *et al.* (1986) provided a simple but efficient methodology for dealing with the problem, but neither their solution nor more complex alternatives, such as the above-mentioned computer programs, take into account many of the details we discuss here. These concern specificity and heterogeneity issues (section 3), penetrance rate estimates from families undergoing molecular testing (section 4), lack of information about the phenotype of the pedigree generator (section 5), genealogies containing grouped parent-offspring information (section 6), and ascertainment issues responsible for the inflation of K estimates (section 7). In section 1 we briefly review the method for estimating the penetrance rate from pedigree data, and in section 2 we make a digression on the determination of the exact credible interval for this estimate.

(1) Method for Estimating the Penetrance Rate K

More details on the method described below are found in the original paper by Rogatko *et al.* (1986) and Horimoto *et al.* (2010). The first step of the method consists in trimming or filtering the pedigree information, that is, replacing the original pedigree with one containing only individuals that are informative or relevant with respect to penetrance estimation. Expressions like trimming and trimmed seem to be more appropriate, but we shall keep the nomenclature coined originally by Rogatko *et al.* (1986). As an example we will consider the hypothetical filtered pedigree shown in Figure 1, with several individuals affected by
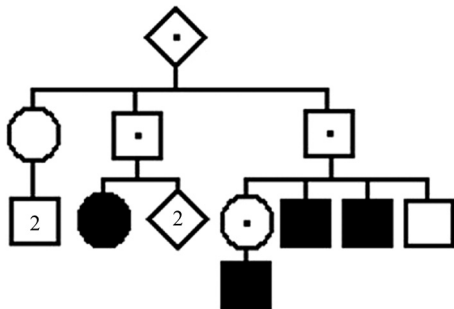
Figure 1 - Filtered pedigree of a family with several cases of an incompletely penetrant autosomal disease.

a rare autosomal dominant condition. The individual of the first generation is the genealogy or pedigree generator. The symbols marked with a point indicate obligate normal (non-penetrant) heterozygous carriers of the gene, and the darkened symbols represent affected (penetrant) heterozygotes.

The filtered pedigree contains four affected individuals, four normal obligate carriers, three normal offspring of obligate carriers, and one tree of normal individuals descendants from an obligate carrier (one normal female with two normal male offspring, shown at the leftmost position of the pedigree). Letting K be the penetrance rate value, the probabilities associated with each of these four different structures are, respectively, $K/2$, $(1-K)$ in the case of the pedigree generator, or $(1-K)/2$ in the case of the other three normal obligate carriers, $(2-K)/2$, and $\{1/2 + (1-K)/2.[(2-K)/2]^2\}$.

The likelihood function, that is the probability of occurrence of the pedigree conditional to the observed structures occurring in it, is derived from the quantities associated with these structures. In the present case, by neglecting constant values unimportant in the maximization procedure that will follow, the likelihood function takes the form $p = K^4(1-K)^4 (2-K)^3[4+(1-K)(2-K)^2]$. By solving the equation $dP/dK = 0$ (or, more conveniently, $dL/dK = d\log(P)/dK = 4.\log(K)+4.\log(1-K)+3.\log(2-K)+\log[4+(1-K)(2-K)^2] = 0$), we obtain the maximum likelihood estimate of the penetrance value K, which for this family takes the value of 0.418.

Heterozygosis probabilities and the corresponding risks for the offspring of all individuals of the filtered pedigree can then be determined without difficulties. Obligate carriers (known non-penetrant carriers and affected penetrant heterozygotes) have genotype Aa and the risk for their offspring is simply $R_1 = K/2 = 0.418/2 = 0.209$, or approximately 21%, for the above shown example. The probability of heterozygosis for normal individuals born to obligate carriers (three of which occur in the family used as example) is taken directly from the quantity $(2-K)/2 = 1/2+(1-K)/2$ as $P(het) = [(1-K)/2]/[(2-K)/2] = (1-K)/(2-K) = 0.582/1.582 = 0.368$. The probability of affected offspring for these individuals is then $R_2 = (1-K)/(2-K).K/2 = 0.368 \times$

$0.209 = 0.077$, or approximately 8%. The heterozygosis probabilities for all three individuals of the single tree of normal individuals occurring in the worked pedigree can be obtained from the term $\{1/2 + (1-K)/2.[(2-K)/2]^2\}$ by applying simple Bayesian reasoning, or by means of computer programs. In an earlier work (Horimoto *et al.*, 2010) we describe two self-contained computer programs that perform most calculations necessary to estimate the penetrance rate. These are the programs *PenCalc for Windows* and *PenCalc Web*, which can be obtained free of charge from the web page *http://www.ib.usp.br/~otto/software.htm*. Both programs are described in detail in the above mentioned article as well as in the PDF-guide included in the zipped file of the program PenCalc for Windows.

(2) Construction of Exact Credible Intervals for K Estimates

Rogatko *et. al*. (1986) also used an exact credible interval associated with a given K estimate. This interval can be obtained by finding the area that corresponds to a given proportion (v.g., 95%) of the total area under the graph of the likelihood function. Mathematically, the problem is reduced to integrating the function $y = f(K)$ between two limits a and b with the same ordinate value $[f(K=a) = f(K=b)]$, so that $\int_{a,b}[f(K)dK]/ \int_{0,1}[f(K)dK] = 0.95$, an operation which can be accomplished by simple computer programs using numerical integration techniques such as Romberg's oscullatory method. The lower and upper limits of the exact 95% credible interval for the estimate $K = 0.418$ of the example above are 0.163 and 0.725, respectively. This credible interval is so large that it might seem to be impractical in a clinical setting. The reason for this particular extreme range is that it was derived from the few data of the small family used as example. In practice, larger pedigrees are usually used. The ideal situation is one where several pedigrees of the same condition are available for analysis, and the pooled data are used to perform the calculations of the penetrance rate and of its 95% credible interval. For instance, from the analysis of 21 different published pedigrees on the autosomal dominant ectrodactyly-tibial hemimelia syndrome, penetrance estimates and their corresponding credible intervals varied from 0.191 (0.044-0.574) to 0.750 (0.329-0.973), while the global (pooled) penetrance value estimate was 0.392, with a 95% credible interval of 0.339 to 0.447 (Horimoto, 2009).

(3) Specificity and Heterogeneity Issues

Another point that merits discussion is whether the K value is specific for the family in which the disease segregates or for the condition itself, independently from the family. The non-penetrance of a genetic trait is assumed to represent the lack of its phenotypic manifestation exclusively or predominantly due to environmental factors (Murphy and Chase, 1975; Praxedes and Otto, 2000) or random

genetic and epigenetic processes linked to the disease locus. Of course penetrance can also be affected by a number of events that include the epistatic action of modifying genes and even temporal modifications of diagnostic criteria. Therefore, to a certain extent, penetrance estimates might be family-specific. Another complicating issue is that genetically heterogeneous conditions can be merged in the pooling process. Nevertheless, since the statistical credible intervals of isolated pedigrees usually are large, pooled estimates of the parameter should be preferred, unless statistical tests disclose the existence of great amounts of heterogeneity among penetrance estimates from various pedigrees.

(4) Penetrance Rate Estimates from Families Undergoing Molecular Testing

In this section we discuss the comparison of estimates obtained from families without molecular testing as to those for which DNA testing has been used for classifying non-penetrant heterozygotes and normal homozygotes. In the latter case, if molecular testing discloses all non-penetrant heterozygotes inside normal trees of individuals descending from obligate carriers, and if there are n_1 affected (penetrant) individuals and n_2 non-penetrant heterozygotes in the family, the likelihood function reduces to $L = \log(P) = n_1.\log(K)+n_2.\log(1-K)$. The maximum likelihood estimate is then $K = n_1/(n_1+n_2)$, with binomial sampling variance of $var(K) = K(1-K)/(n_1+n_2)$. This would be an ideal situation in which, besides providing a better estimate of K, the corresponding 95% credible interval of the penetrance value thus evaluated will be much smaller than the one provided by the analysis of the family without DNA testing.

(5) Lack of Information about the Phenotype of the Pedigree Generator

In some published pedigrees there is a lack of phenotypic information about the genealogy generator (affected or non-affected?). Furthermore, the likelihood function P may not include the parameter K or $(1-K)$ corresponding to the genealogy generator.

In order to evaluate whether the inclusion of the pedigree generator significantly alters this K estimate, it is not necessary to repeat the calculations for the two configurations possible (penetrant or non-penetrant common ascendant), because the likelihood function P, derived without information on the pedigree generator, is correct, and thus cannot be improved. In fact, if one wants to refer to the pedigree founder, one can say that she/he was affected with probability K and unaffected with probability $(1-K)$. The resulting likelihood is $KP + (1-K)P = P$.

(6) Genealogies Containing Grouped Parent-Offspring Information within Trees of Normal Individuals Descending from Obligate Carriers

Certain published pedigrees present grouped parent-offspring trees of normal individuals, without informing the corresponding offspring numbers of all individuals in a given sibship, as is the case of the pedigree with cases of the ectrodactyly-tibial hemimelia syndrome (Majewski *et al.*, 1985) shown in Figure 2. This tree of normal individuals represented by individuals II.8 to II.11 and III.14 to III.28 does not detail individual offspring numbers, and only the total number of 15 is given.

Incomplete pedigree information is a simple but interesting problem in combinatorial analysis that can be straightforwardly solved by means of the theory of difference operators. Table 1 lists the numbers of possible genealogy structures for a case of incomplete parent-offspring information as a function of both parent and offspring numbers. Therein, with four parents, the number of possible structures is given by $y_4(n) = (n+1)^2 + (n+1)n(n-1)/6$, where n is the total offspring number of the four parents. For $n = 15$, the outcome is $y_4(15) = 816$ of such structures.

For each combination $\{i, j, l, m\}$ of offspring number the likelihood function of the whole tree is:

$$K^4(1-K)^2(2-K)^5\{1/2+(1-K)/2[(2-K)/2]^9\}\{1/2+(1-K)/2.[(2-K)/2]^i\}\{1/2+(1-K)/2.[(2-K)/2]^j\}\{1/2+(1-K)/2.[(2-K)/2]^l\}\{1/2+(1-K)/2.[(2-K)/2]^m\},$$

where i, j, l, and m are the unknown numbers of children for each of the individuals II-8, II-9, II-10, and II-11, respec-

Table 1 - Numbers of possible genealogy structures for the case of incomplete parent-offspring information.

Offspring number	Parent number							
	1	2	3	4	5	6	7	8
0	1	1	1	1	1	1	1	1
1	1	2	3	4	5	6	7	8
2	1	3	6	10	15	21	28	36
3	1	4	10	20	35	56	84	120
4	1	5	15	35	70	126	210	330
5	1	6	21	56	126	252	462	792
6	1	7	28	84	210	462	924	1716
7	1	8	36	120	330	792	1716	3432
8	1	9	45	165	495	1287	3003	6435
9	1	10	55	220	715	2002	5005	11440
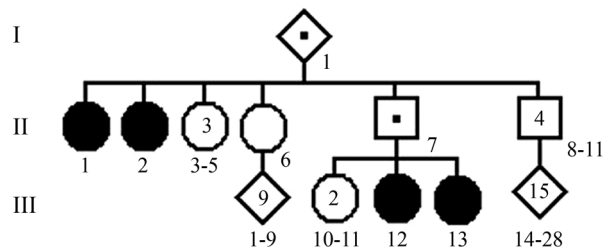
Figure 2 - Pedigree containing a grouped tree of normal individuals (modified from Majewski *et al.*, 1985).

tively. In a population of approximately stable size, the average offspring number per couple does not differ from two and it is known that the number of children per couple adequately fits a Poisson distribution $g(x) = e^{-2}2^x/x!$.

If each possible configuration is weighed by its probability (according to the Poisson distribution for the number of children per couple), this gives a credible interval on K in a straightforward manner. This can be achieved as follows using the function:

$P(i,j,l,m) = K^4(1-K)^2(2-K)^5 \{1/2+(1-K)/2[(2-K)/2]^9\}$
$\{1/2+(1-K)/2.[(2-K)/2]^i\}\ e^{-2}2^i/i!\{1/2+(1-K)/2.$
$[(2-K)/2]^j\}e^{-2}2^j/j!\{1/2+(1-K)/2[(2-K)/2]^l\}e^{-2}2^l/l!\{1/2$
$+(1-K)/2[(2-K)/2]^m\}e^{-2}2^m/m!,$

and considering all 816 possible configurations referred to above. For each configuration one can then obtain not only a K_{ijlm} estimate but also its exact 95% credible interval. Estimates for the penetrance value K, as well as for its exact 95% lower and upper credible limits $_LK$ and $_UK$, corresponding to the given tree structure can be straightforwardly obtained by averaging the estimates K_{ijlm}, as well as those for the lower and upper limits $_LK_{ijlm}$ and $_UK_{ijlm}$, as:

$K = \Sigma(K_{ijlm}.e^{-2}2^i/i!.e^{-2}2^j/j!.e^{-2}2^l/l!.e^{-2}2^m/m!) / \Sigma(e^{-2}2^i/i!$
$. e^{-2}2^j/j! . e^{-2}2^l/l! . e^{-2}2^m/m!) = \Sigma[K_{ijlm}. 2^{i+j+l+m}/(i! . j! .$
$l! . m!)] / \Sigma[2^{i+j+l+m}/(i! . j! . l! . m!)] = \Sigma[K_{ijlm}. 1/(i! .$
$j! . l! . m!)] / \Sigma[1/(i! . j! . l! . m!)] .$

The numerical procedure is herein detailed using as an example the simple hypothetical pedigree represented in Figure 3. Table 2 shows the penetrance rate estimates for all possible configurations contained in the grouped tree of normal individuals of Figure 3. The final estimates for the penetrance rate and for the lower and upper limits of its 95% credible interval are 0.4377, 0.1471 and 0.7813, respectively.
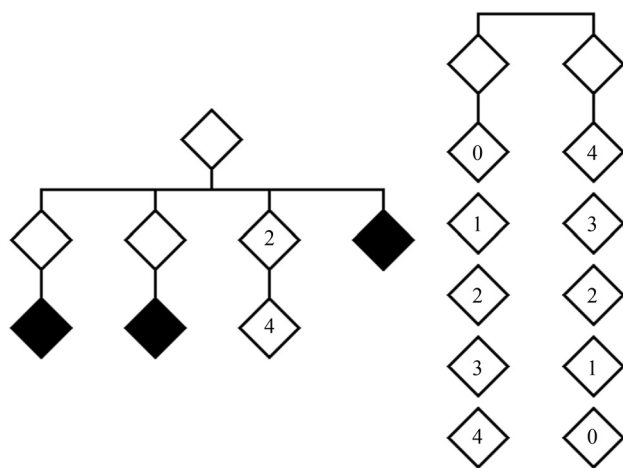
Figure 3 - Hypothetical pedigree containing a grouped tree of normal individuals with incomplete information (at left) and the possible configurations obtained from the grouped tree (at right).

Table 2 - Maximum likelihood estimates of penetrance value (K_{ij}) and lower ($_LK_{ij}$) and upper ($_UK_{ij}$) limits of its correspondent 95% credibility interval for all possible configurations obtained from the grouped tree of normal individuals shown in the pedigree of Figure 3.

i	j	P_{ij}	K_{ij}	$_LK_{ij}$	$_UK_{ij}$
0	4	0.0625	0.443106	0.149314	0.782835
1	3	0.2500	0.437653	0.147112	0.781286
2	2	0.3750	0.435844	0.146389	0.780826
3	1	0.2500	0.437653	0.147112	0.781286
4	0	0.0625	0.443106	0.149314	0.782835
		1.0000	0.437656	0.147116	0.781307

i, j: offspring numbers of the two parents; P_{ij}: normalized product (weighing factor) obtained through $P_{ij} = e^{-2}2^i/i!.e^{-2}2^j/j!/\Sigma(e^{-2}2^i/i!.e^{-2}2^j/j!)$; average estimates for K_{ij}, $_LK_{ij}$ and $_UK_{ij}$ are shown in bottom line.

(7) Ascertainment Issues

The general method proposed by Rogatko *et al.* (1986) does not take into account any ascertainment biases. The authors are correct in their paper in stating that their approach gets around the sample space problem by using only the likelihood of the parameters, given the actual observations. Yet if ascertainment is not included, that likelihood itself will not be correct. Advanced computer programs that perform segregation analysis or estimate linkage, such as the classical S.A.G.E. (S.A.G.E., 2009) and LINKAGE (Lathrop *et al.*, 1985) programs we referred to in the introduction section, do not apply any ascertainment bias to the penetrance rate they indirectly estimate.

By using a very simple example we could show that the crude K estimates obtained from genealogies are actually inflated. Figure 4 lists all possible trees of offspring size = 2 with a pedigree generator carrier of the pathologic gene (affected in A, B and C, and non-penetrant heterozygous in D, E and F) disclosed by an (impossible) ascertainment devoid of any bias. The probabilities associated with each tree are shown in Figure 4.

Let now n_A, n_B, n_C, n_D, n_E, and n_F be the numbers of structures A, B, C, D, E, and F observed in an ideal, large sample collected without any ascertainment bias. Then, the corresponding likelihood function in logarithmic form would be:

$L_1 = (3n_A+2n_B+2n_D+n_E).\ln(K) + (n_D+n_E+n_F).\ln(1-K)$
$+ (n_B+2n_C+n_E+2n_F).\ln(2-K),$

from which the maximum likelihood estimate of K is obtained without difficulties.

A careful collection of a large number of families with offspring number 2 and a tree-generator carrier of the gene would consist only of structures A, B, C and D. Configuration E would not be included, as the only affected individual would be, with a large probability, the result of a new mutation; and configuration F would never be ascer-
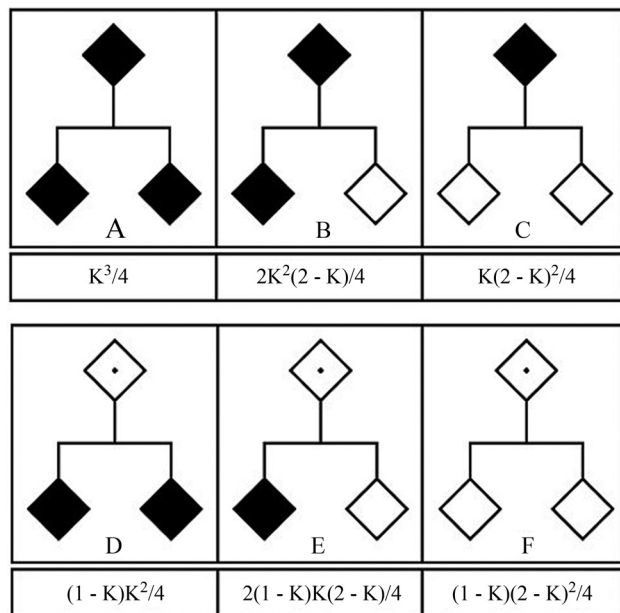
Figure 4 - Probabilities associated with all possible trees of offspring size = 2, in which the pedigree generator is an obligate carrier of a pathologic gene. The pedigree generator is affected in the first three trees and unaffected in the remaining ones.

A	B	C
$K^3/4$	$2K^2(2-K)/4$	$K(2-K)^2/4$
D	E	F
$(1-K)K^2/4$	$2(1-K)K(2-K)/4$	$(1-K)(2-K)^2/4$

tained, because it contains only normal individuals. The corresponding (logarithmic) likelihood expression would then be given by:

$$L_2 = (3n_A+2n_B+n_C+2n_D).\ln(K) + n_D.\ln(1-K) + (n_B+2n_C).\ln(2-K),$$

from which, as in the previous case, the maximum likelihood estimate can be easily obtained.

Let us now take the following numerical example. Let the actual (unknown) value of K be 0.8; then the probabilities associated with structures A, B, C, D, E, and F would take the values P(A) = 0.128, P(B) = 0.384, P(C) = 0.288, P(D) = 0.032, P(E) = 0.096, and P(F) = 0.288. In a sample of size 1000 we would therefore expect to find the sample numbers $n_A = 128$, $n_B = 384$, $n_C = 288$, $n_D = 32$, $n_E = 96$, and $n_F = 72$. The unbiased estimate would then take the value $K = 0.8$, as expected. In the case of an incompletely ascertained sample, the biased estimate of K' would take the value 0.951825 > 0.8.

It is not possible to obtain an exact solution in simple analytical form for the function $K' = f(K)$, where K' is the biased maximum likelihood estimate and K the true one (unknown, completely unbiased estimate of the penetrance value), but we can evaluate K' estimates for any given fixed value of K by means of likelihood expression L_2. For any true value of K the biased estimate, K' is an inflated value, as we could guess intuitively. Using a program on non-linear regression analysis, such as the NLREG software (Sherrod, 2000), we can adjust the observed set of points to the generalized empirical function $y = ax^b.e^{cx}$ (Bronshtein and Semendiaev, 1973) where $y = K'$ and $x = K$.

We then estimated sets of pairs of values K and K' varying offspring size from 2 to 10, and in each case we obtained corresponding generalized empirical functions $y_i = a_i x^{bi}.e^{cix}$, where $y = K'$ and $x = K$. as in the case of the previous example. The functions corresponding to offspring sizes from 2 to 10, all showing a perfect statistical fitting with the corresponding observed biased estimates, are plotted in Figure 5, where y stands for K' and x for K. The graph also shows the function $K' = K$ that corresponds to the case of an offspring with infinite size.

As expected, with sibship size increasing the difference between corresponding estimates K' and K becomes negligible, mainly in relation to values of K in its usual range ($K > 0.8$). This is a result that certainly can be generalized for any homogeneous or heterogeneous set of
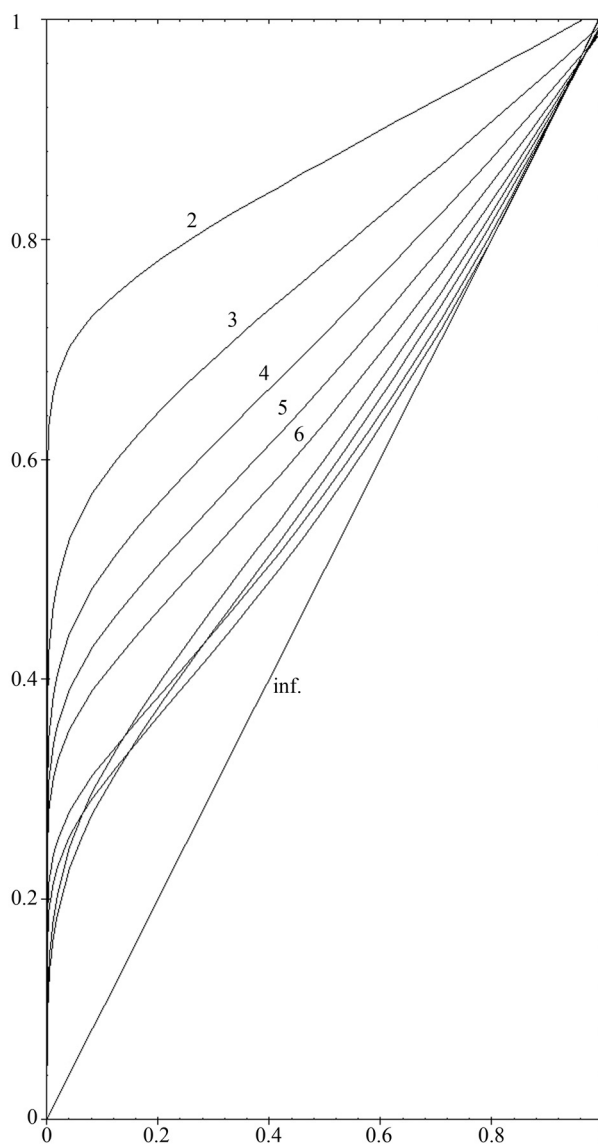
Figure 5 - Relation between unbiased (K) and biased (K') penetrance values, shown at abscissa and ordinate axes respectively, depending on offspring size (2, 3, 4, 5, 6, ..., infinite).

pedigrees. Since optimized K estimates are obtained from large filtered pedigrees, or from the pooling of many pedigrees, the ascertainment bias just discussed will only produce slightly inflated K estimates. In the case the actual values of K, as well as the total number of informative individuals (penetrant, obligate non-penetrant and those belonging to normal trees descending from obligate carriers) are both small, the K estimates will not be reliable, as shown in Figure 5. For offspring sizes of 10 (total of 11 informative individuals) or more, it is also easy to conclude that estimated values of K in the range of 0.5 or more are reliable and do not need to be corrected. In any case, estimate corrections can be performed by enumerating all possible filtered pedigrees corresponding to a given tree structure and comparing the estimated K values to the inferred actual ones, just as we did before using the very simple examples discussed above.

References

Bronshtein I and Semendiaev K (1973) Manual de Matemáticas para Ingenieros y Estudiantes. 2nd edition. Editorial Mir, Moscow, 694 pp.

Horimoto ARVR, Onodera MT and Otto PA (2010) PENCALC: A program for penetrance estimation in autosomal dominant diseases. Genet Mol Biol 33:455-459.

Lathrop GM, Lalouel JM, Julier C and Ott J (1985) Multilocus linkage analysis in humans: Detection of linkage and estimation of recombination. Am J Hum Genet 37:482-498.

Majewski F, Küster W, Haar B and Goecke T (1985) Aplasia of tibia with split-hand/split-foot deformity. Report of six families with 35 cases and considerations about variability and penetrance. Hum Genet 70:136-147.

Murphy EA and Chase GA (1975) Principles of Genetic Counseling. Year Book Medical Publishers, Chicago, 391 pp.

Praxedes LA and Otto PA (2000) Estimation of penetrance from twin data. Twin Res 3:294-298.

Rogatko A, Pereira CAB and Frota-Pessoa O (1986) A Bayesian method for the estimation of penetrance: application to mandibulofacial and frontonasal dysostoses. Am J Med Genet 24:231-246.

Internet Resources

Horimoto ARVR (2009) Estimativa do valor da taxa de penetrância em doenças autossômicas dominantes: Estudo teórico de modelos e desenvolvimento de um programa computacional (Penetrance rate estimation in autosomal dominant diseases: theoretical study and development of a computer program). PhD Thesis, Instituto de Biociências, Universidade de São Paulo, São Paulo, 2009. http://www.teses.usp.br/teses/disponiveis/41/41131/tde-15102009-161545/ (April 16, 2012).

Horimoto ARVR and Otto PA (2008) Penetrance rate estimation. In: Otto PA (ed) Métodos Clásicos y Modernos para el Análisis de Datos en Genética Humana. EdUNaM, Posadas, pp 123-150, http://www.ib.usp.br/~otto/pop_genetics.htm or http://www.lacygh.com.ar/abajo.htm (April 16, 2012).

S.A.G.E. (2009) Statistical Analysis from Genetic Epidemiology. Release 6.0.1, http://darwin.cwru.edu/ (April 16, 2012).

Sherrod PH (2000) Non Linear Regression Analysis Program (NLREG), http://www.nlreg.com/ (April 16, 2012).

Associate Editor: Angela M. Vianna-Morgante